From r to Q^*: Your Language Model is Secretly a Q-Function

Year: 2024 | Citations: 215 | Authors: Rafael Rafailov, Joey Hejna, Ryan Park, Chelsea Finn

---

Abstract

Reinforcement Learning From Human Feedback (RLHF) has been critical to the success of the latest generation of generative AI models. In response to the complex nature of the classical RLHF pipeline, direct alignment algorithms such as Direct Preference Optimization (DPO) have emerged as an alternative approach. Although DPO solves the same objective as the standard RLHF setup, there is a mismatch between the two approaches. Standard RLHF deploys reinforcement learning in a specific token-level MDP, while DPO is derived as a bandit problem in which the whole response of the model is treated as a single arm. In this work we rectify this difference. We theoretically show that we can derive DPO in the token-level MDP as a general inverse Q-learning algorithm, which satisfies the Bellman equation. Using our theoretical results, we provide three concrete empirical insights. First, we show that because of its token level interpretation, DPO is able to perform some type of credit assignment. Next, we prove that under the token level formulation, classical search-based algorithms, such as MCTS, which have recently been applied to the language generation space, are equivalent to likelihood-based search on a DPO policy. Empirically we show that a simple beam search yields meaningful improvement over the base DPO policy. Finally, we show how the choice of reference policy causes implicit rewards to decline during training. We conclude by discussing applications of our work, including information elicitation in multi-turn dialogue, reasoning, agentic applications and end-to-end training of multi-model systems.